BioGecko

Vol 12 Issue 03 2023
ISSN NO: 2230-5807

TEXT SUMMARIZATION ON YOUTUBE VIDEOS IN EDUCATIONAL DOMAIN

Amit Bhagat
Computer Technology Department
Yeshwantrao Chavan College of Engineering
Nagpur, India
amitbhagat963@gmail.com

MrunalSingade
Computer Technology Department
Yeshwantrao Chavan College of Engineering
Nagpur, India
mrunalsingade@gmail.com

Prasanna Anjankar
Computer Technology Department
Yeshwantrao Chavan College of Engineering
Nagpur, India
prasannanjankar@gmail.com

SavinaySurbhik
Computer Technology Department
Yeshwantrao Chavan College of Engineering
Nagpur, India
savinaysurbhik@gmail.com

NileshSambhe
Computer Technology Department
Yeshwantrao Chavan College of Engineering
Nagpur, India
nilesh.sambhe@gmail.com

***Abstract*—**Text Summarization is the process of converting a text into a compressed formwhile the significant information is still maintained and the meaning of the text remains the same. Automatic text summarization help us to findrelevant information in large text documents in fast and efficient manner with very less or no effort. This method suggests an application for creating summary of the youtube video transcripts based on some natural language processing (NLP) algorithms without distorting the actual meaning of the text. The Onlineeducational system is becoming more and more popular since the start ofCovid-19 and it shows no sign of decline. This rise of online educationsystem is also witnessing numerous hours of educational videos gettinguploaded on the platforms like Youtube daily. There is also a term “click-bait” which is mainly associated with misleading title or thumbnail. Thesystem mainly aims at the educational videos. In searching a particularvideo that contains the information of our interest, a significant amount oftime is wasted in watching all the similar yet unimportant videos. In the paper the developed application aims to decrease the transcript size and eventually the time of theuser spent on finding a video of their interest. It also gives users sixdifferent algorithms to choose and summarize the transcript. Thisapplication first retrieves the provided or automatically generated subtitles of the video fromyoutube_transcript_api available in

python using the link of the video enteredby the user. The application takes video link, summary algorithm andsummary percentage as input and provides the summary of the transcriptas output.

Keywords—transcript, text, algorithm, summarization.

I. INTRODUCTION

The Online educational system is becoming more and more popular since the start of Covid-19 and it shows no sign of decline. This rise of onlineeducation system is also witnessing numerous hours of educational video getting uploaded on platforms like Youtube daily. According to the IPLIX Content Consumption survey, 52.5% of those surveyed say they use social media on a daily basis for 2-4 hours . It is frustrating and time consuming to watch a (say) 15 minute video only to find out that it doesn't contain what we intend to find. There is also a term "click-bait" which ismainly associated with misleading title or thumbnail. Also, Users like students who prepare for anexamination by watching such videos cannot go through whole video in shorter time.

The Youtube Transcript Summarizer will reduce the transcript and provide the summary of the transcript. This will help the user to find the desired youtube video without wasting time watching all the potentialvideos. In addition it will also provide multiple language translation support. Text Summarization is the process of converting a text into a compressed form while the significant information is still maintained and the meaning of the text remains the same. Automatic text summarization help us to find relevant information in large text documents in fast and efficient manner. The summarization approaches can broadly classified into two types, 1. Extractive text summarization 2. Abstractive text summarization

In this web application, we will be using Extractive Text summarization approach.

II. PROPOSED METHODOLOGY

A. Mathematical Foundations

TF-IDF (Term frequency-inverse document frequency) is a numeric value that is employed in the disciplines of information retrieval (IR).

TF(Term frequency)

It works by calculating the number of occurrences of a specific term in relation tothe document.

IDF(Inverse document frequency) examines how common (or rare) a word is in the corpus.

IDF is calculated as follows:

Where:

- t : term (word) under examination,
- N is the total nuumber of documents (d) in the corpus (D)
- Number of documents that contain the term(t) is the denominator.

Scikit-Learn

- $IDF(t) = \log(1+n/1+df(c))+1$

Standard Notation

- $IDF(T)=\log(n/df(t))$

Combining both gives TF-IDF.

The basic idea which is suggested by TF-IDF is that the importance of a term is inversely related to its frequency across documents.

$Tfidf(t,d,D)=tf(t,d).idf(t,d)$

B. Available Algorithms

User can use six algorithm to summarize the transcript.

1) Text Rank algorithm based(sumy)

With keyword extractions from documents, Text Rank is a graph-based summarization method.

```
from sumy.summarizers.text_rank import TextRankSummarizer
summarize_TR_sumy = TextRankSummarizer()
summ_TR_sumy =summarize_TR_sumy(parser.document,2)
for sentence in summ_TR_sumy:
print(sentence)
```

2) Luhn algorithm base (sumy)

A well-known IBM researcher who gave it its name proposed this algorithm. It evaluates the sentences on the basis of the frequency of the most important words.

```
fromsumy.summarizers.luhn import LuhnSummarizer
summarizer_Luhn_sumy_ = LuhnSummarizer()
summ_Luhn_sumy =summarizer_Luhn_sumy(parser.document,2)
for sentence in summa_Luhn_sumy:
 print(sentence)
```

3) Text Rank Algorithm based (gensim)

This algorithm works on the basis of PageRank algorithm for ranking search results.

1. It Pre-process the text provided. Punctuations, stop words, and stemming are included in this.
2. Creates a graph using sentences that are the vertices.
3. The edges of the graph which represents the similarity between two sentences at the vertices.
4. Run the PageRank algorithm on the weighted graph.
5. Take out vertices with the highest scores and add them to the summarized text.
6. Count of vertices to be considered are decided on the basis of word count.

4) Latent Semantic analysis based (sumy)

Unsupervised text summarization using latent semantic analysis combines term frequency approaches with singular value decomposition. This technique is one of the latest suggested methods for summerization

```
from sumy.summarizers.lsa import LsaSummarizer
summarizer_LSA_sumy= LsaSummarizer()
summary_LSA_sumy=summarizer_LSA_sumy(parser.document,2)for sentence in summary_2:
 print(sentence)
```

5) Frequency based(NLTK)

Working of NLTK algorithm is based on distribution of the frequency of words. Frequency distribution of words is generated with the use of FreqDist () function in NLTK.

```
fromnltk.probability import FreqDist
frequency_distribution_of_words = FreqDist(tokenized_word)
print(frequency_distribution_of_words)
```

6) Frequency based(spacy)

Written in the Python and Cython programming languages, SpaCy is an open-source library which is used for sophisticated NLP. Under the MIT licence, the library is distributed.

C. Use case diagram

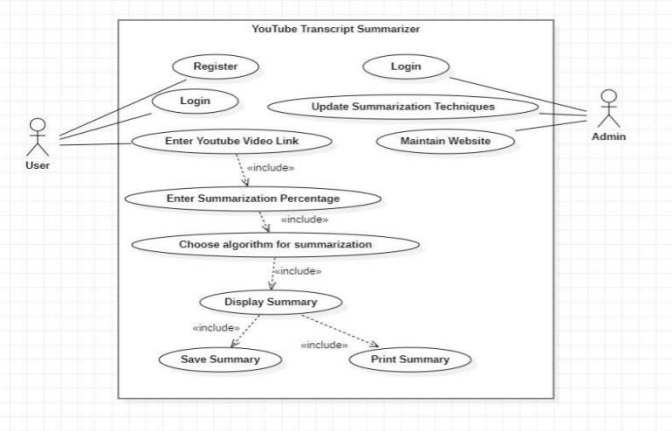

Figure2.1: Use Case Diagram

- User should enter the youtube link for which the summary has to be taken out.
- The user will select the percentage of summary and the Algorithm for summarization.
- The summary will be displayed on the screen. The user can either select to choose important points.
- User can save or print the summary as perrequirement.
- The admin will take care about Login, Updating the summary algorithms and maintain the website.

D. ACTIVITY DIAGRAM

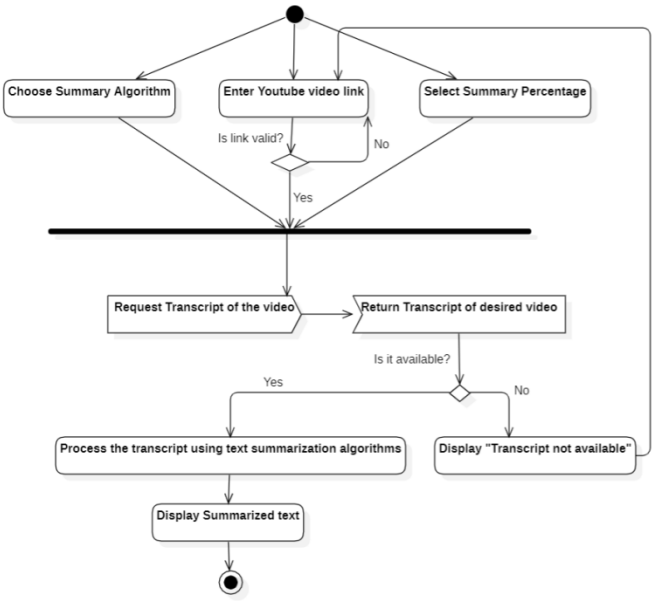

Figure 2.2 Activity Diagram

- First the user has to choose the Summary Algorithm Choose the summary percentage. Then, Enter the Youtube video link.
- The algorithm will fetch the transcript if the transcript is available and will use the auto generated transcript if its not provided by the uploader.
- If both the conditions are false i.e., neither the subtitles are available more the auto generate subtitle option is enabled by the author of video while uploading then it will return "No transcript available".

III. SOFTWARE IMPLEMENTATION

A. The interface Home page of the web app :

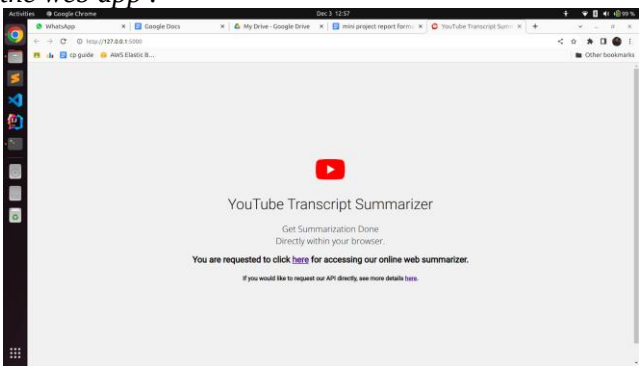

Figure 3.1: Interface

Click on the "here" text to start summarizing the desired youtube video.

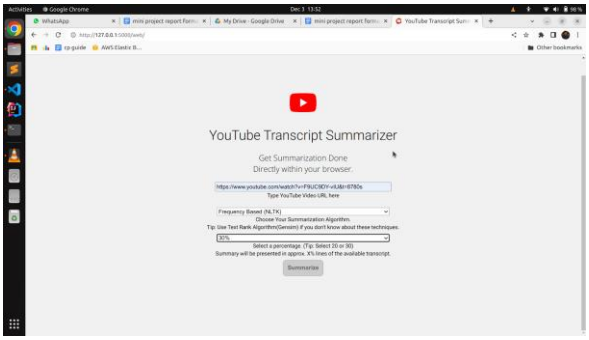

Figure 3.2: Opening screen

Here we can choose the algorithm for summarizing as well as the percentage upto which the text has to be summarized.

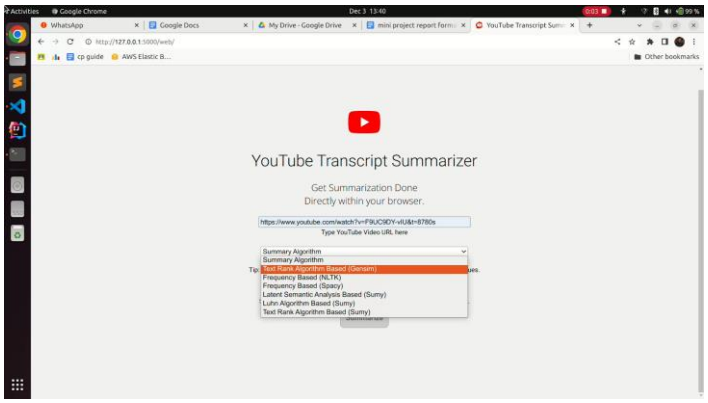

Figure 3.3: Selection of Algorithm.

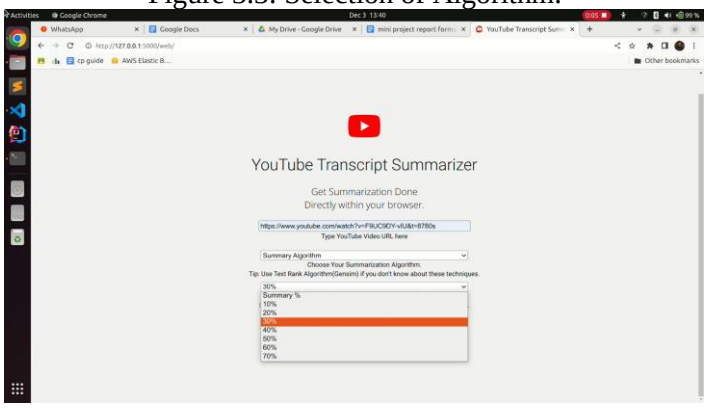

Figure 3.4: Selection of Percentage of summary.

After clicking on “summarize” button, The summarized text will appear in a moment.

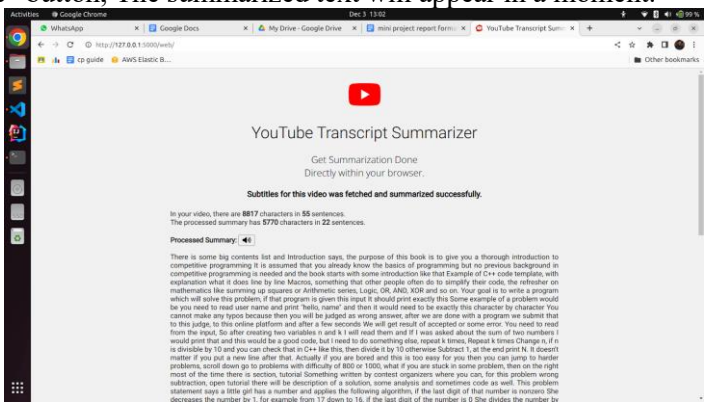

Figure 3.5: Summarized Text.

IV. RESULT AND DISCUSSION

The final application gives the user six summary algorithms to choose and summarize their youtube video transcript. It has been noted that Text Rank algorithm based(sumy) provides the shortest summary for the same video and same summary percentage. Luhn algorithm base (sumy) provides the second shortest summary. Text Rank Algorithm based (gensim) and Latent Semantic analysis based (sumy) provides the exact same number of characters but the selection of words are different.Frequency based(NLTK) and Frequency based(spacy) comes at fifth and sixth position respectively. Finally, the application provides user with variety of algorithms to choose to summarize their video transcript.

V. Conclusion and Future Scope

Our project proposes solution to summarize transcripts of lengthy youtube videos. In our project we have used several algorithms to effectively summarize the transcript. It gives users six different algorithms to choose and summarize the transcript. This application first retrieves the subtitles/transcript provided by the uploader or the automatic generated subtitles/transcript of the video from youtube_transcript_api available in python using the video link entered by the user. The application takes video link, summary algorithm and summary percentage as input and provides the summary of the transcript as output. This solution works only if the transcript is provided by the author or auto-generation is allowed while uploading the video, the project can be expanded by improving the application such that it work for the videos without transcript or auto-generation allowed. Multi language support can also be added so that transcript of educational videos in multiple languages can be supported.

VI. Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] A. N. S. S. Vybhavi, L. V. Saroja, J. Duvvuru and J. Bayana, "Video Transcript Summarizer," 2022 International Mobile and Embedded Technology Conference (MECON), 2022, pp. 461-465, doi: 10.1109/MECON53876.2022.9751991.

[2] K. Kulkarni and R. Padaki, "Video Based Transcript Summarizer for Online Courses using Natural Language Processing," 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2021, pp. 1-5, doi: 10.1109/CSITSS54238.2021.9683609.

[3] L. Herranz and J. M. Martínez, "A Framework for Scalable Summarization of Video," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, no. 9, pp. 1265-1270, Sept. 2010, doi: 10.1109/TCSVT.2010.2057020.

[4] HarikaUpganlawar, NileshSambhe, "Surveillance of suspicious discussions on online forums using text data mining," International Journal of Advances in Electronics and Computer Science, vol. 4, issue 4 ,pp. 50-51, April 2017,

[5] KedarBellare, Anish Das Sarma, Atish Das Sarma, NavneetLoiwal, Vaibhav Mehta, Ganesh Ramakrishnan, and Pushpak Bhattacharyya. 2004. “Generic Text Summarization Using WordNet”. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04), Lisbon, Portugal.European Language Resources Association (ELRA).

[6] C. M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon and E. J. Delp, "Automated video program summarization using speech transcripts," in IEEE Transactions on Multimedia, vol. 8, no. 4, pp. 775-791, Aug. 2006, doi: 10.1109/TMM.2006.876282.

[7] Rahul, S. Adhikari and Monika, "NLP based Machine Learning Approaches for Text Summarization," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 535-538, doi: 10.1109/ICCMC48092.2020.ICCMC-00099.

[8] Christian, Hans & Agus, Mikhael & Suhartono, Derwin. (2016). Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF). ComTech: Computer, Mathematics and Engineering Applications. 7. 285. 10.21512/comtech.v7i4.3746.

[9] Rahul, S. Adhikari and Monika, "NLP based Machine Learning Approaches for Text Summarization," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 535-538, doi: 10.1109/ICCMC48092.2020.ICCMC-00099.

[10] A. Dilawari and M. U. G. Khan, "ASoVS: Abstractive Summarization of Video Sequences," in IEEE Access, vol. 7, pp. 29253-29263, 2019, doi: 10.1109/ACCESS.2019.2902507.

[11] KamelSmaïli, Dominique Fohr, Carlos González-Gallardo, Michal Grega, LucjanJanowski, et al. A First Summarization System of a Video in a Target Language. Springer Proceedings MISSI 2018, 2018. <hal-01819720>

[12] Sah, Shagan& Kulhare, Sourabh & Gray, Allison & Venugopalan, Subhashini & Prud'hommeaux, Emily & Ptucha, Raymond. (2017). Semantic Text Summarization of Long Videos. 989-997. 10.1109/WACV.2017.115.

[13] Yu-Chyeh Wu, Yue-Shi Lee and Chia-Hui Chang, "VSUM: summarizing from videos," IEEE Sixth International Symposium on Multimedia Software Engineering, 2004, pp. 302-309, doi: 10.1109/MMSE.2004.90.
[14] M. Chandra, V. Gupta and S. K. Paul, "A Statistical Approach for Automatic Text Summarization by Extraction," 2011 International Conference on Communication Systems and Network Technologies, 2011, pp. 268-271, doi: 10.1109/CSNT.2011.65.
[15] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.
[16] https://www.medianews4u.com/87-5-of-indians-spend-most-of-their-time-on-youtube-followed-by-instagram-reveals-iplix-content-consumption-survey/
[17] https://www.irjet.net/archives/V8/i8/IRJET-V8I8411.pdf
[18]https://www.sciencedirect.com/science/article/abs/pii/S095741742030500
[19] Assessing sentence scoring techniques for extractive text summarization: https://booksc.eu/book/21495570/83518a
[20] R. Kannan, G. Ghinea, S. Swaminathan and S. Kannaiyan, "Improving video summarization based on user preferences," 2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013, pp. 1-4, doi: 10.1109/NCVPRIPG.2013.6776187.